The Implementation of Paired Descriptor Functions to Improve Quantitative Structure Activity Relationship Models for Drug Discovery

Aditya Gudibanda, Edward W. Lowe Jr., and Jens Meiler

KEYWORDS. Modeling, QSAR, descriptors

BRIEF. Here we present an improvement in the mathematical description of the constitution and configuration of molecules.

ABSTRACT. One of the major challenges that we face today is the discovery of new and better therapeutics. Pharmaceutical companies and academic institutions screen libraries of thousands of compounds for possible drug candidates in a process called high throughput screening (HTS). This method runs assays on each compound in the library to see which ones have a desired biological activity that might be beneficial for treatment of a disease. However, HTS is cost- and time- intensive [1]. To improve efficiency of HTS, methods have been developed in the field of computational chemistry, which predict the activity of a molecule based on its structure using Quantitative Structure Activity Relationships (QSAR). A machine learning model is trained to understand the complex relationship between molecular structure and biological activity. Eventually it can predict the activity of untested molecules. Previous research has demonstrated the success of this approach leading to the discovery of new biologically active compounds. The algorithms used in this approach utilize descriptor functions, which are tools that mathematically describe the constitution, configuration, and conformation of molecules. Here we present a novel class of descriptor functions that significantly improves the performance of these QSAR models.

INTRODUCTION.

As we move into a new century, we increasingly rely on technology to solve many of the world's major problems. One of these problems is the discovery of not only novel drugs to treat rare or neglected diseases, but also drugs for major diseases that are more efficient than those currently on the market. This is presently accomplished through HTS, the brute force screening of all compounds in a chemical library utilizing robotics in an attempt to find the few that have a desired biological activity, such as the ability to bind to a certain protein. However, this method can be inefficient if molecules are not prioritized. Cost increases linearly with the number of compounds screened making HTS time- and resource-intense.

To address this problem, methods have been developed in the field of computational chemistry, which predict structure-activity relationships. These Quantitative Structure Activity Relationship (QSAR) models correlate the non-linear, highly complex relationship between the structural properties of a molecule with its biological activity.

For example, research by Mueller *et al.* [1] has shown that this process is not only feasible, but also a powerful method that has the capability of discovering active molecules with a novel structure. Thus, these machine learning techniques can find molecules that were previously not considered to possess the desired activity.

A critical factor for achieving efficiency of QSAR models is the ability to encode the constitution, configuration, and conformation of molecules mathematically leveraging "descriptor functions". An improvement of molecule descriptors is one important avenue to improve QSAR models and therefore a major area of research in the field of computational chemistry [1–4].

The goal of this project is to improve the mathematical description of molecules through the implementation of a novel class of molecular descriptors. This new class of descriptors will allow for the encoding of more diverse molecular relationships, such as how different properties among separate regions of a molecule correlate with the observed biological activity. This can be critical when determining whether or not a molecule will bind to a target. We have improved upon the descriptor functions used traditionally by creating versions that have the ability to mix multiple properties. These "paired property descriptors" have significantly greater descriptive power, which allows for greater prediction accuracy for QSAR models. This ultimately aids the process of drug discovery.

MATERIALS AND METHODS.

Dataset.

In order to determine the usefulness of the new descriptor function class, a biologically significant HTS dataset was selected from the publically available PubChem database [5].This dataset had a unique target protein, for which the biologically active and biologically inactive molecules within the dataset had been experimentally determined. The dataset chosen was Assay ID (AID) 891, containing 9379 compounds, of which 1623 are active, 6335 are inactive, and 1640 are inconclusive. The target protein was Cytochrome P450 2D6, which is involved in the metabolism and biotransformation of xenobiotics. It is important in drug discovery primarily for its role in drug metabolism [4], [6], [7].

Novel Paired Descriptor Functions.

a) 2D Autocorrelation

2D Autocorrelation is a descriptor function which uses the topological distance (or spatial lag) between pairwise atoms, and represents these distances with the product of two properties. This method represents the molecule as a graph with atoms at the vertices and bonds as the edges. The topological distance between a pair of atoms is then defined as the smallest number of interconnecting bonds between the two atoms [5].

The 2D Autocorrelation function then sums the property products of all pairs of atoms located at a certain topological distance from each other and returns this value as the element of the vector with the index corresponding to the topological distance. 2D Autocorrelation measures not only the relationships between properties, but also the strength of these relationships. The vector generated by this descriptor category bins the topological distances into one-bond intervals, where the vector element with the index contains , where and are the atom property values of atoms and respectively, and is the Dirac-delta function given by

Equation 1

$$\partial\left(d_n, d_{ij}\right) = \left\{\frac{0, \text{if } d_n \neq d_{ij}}{1, \text{if } d_n = d_{ij}}\right\}$$

Thus, only the atom pairs of the desired topological distance are included in the property product sum in that element of the resulting vector [8], [9].

b) 3D Autocorrelation

3D Autocorrelation is similar to its 2D counterpart, with the major difference being that topological distance is replaced with Euclidean distance. A parameter is chosen for the length of the distance bins to be used in the vector, and each value in the vector is divided by the number of unique distances in that bin. Thus, the vector element in the 3D Autocorrelation vector with index is given by Equation 2.

Equation 2 $$A(d_n) = \frac{1}{2L_n}\sum_{i,j,i\neq j} p_i p_j$$

Here, and are defined similarly as in 2D Autocorrelation, and is the number of unique distance occurring in the distance interval defined by . As this is a measure of continuous rather than discrete data, 3D Autocorrelation is an improvement over 2D Autocorrelation. Additionally, 3D Autocorrelation has the ability to encode the physicochemical distribution of properties within the molecule [10].

c) Radial Distribution Function (RDF)

The highest resolution descriptor considered in this study is the property-weighted Radial Distribution Function (RDF). RDFs weight the property products of pairs of atoms with a Gaussian probability distribution function. RDFs represent the shape and structure of a molecule by combining the frequencies of all pairwise distances between atoms and weighting the property products of all pairs of atoms and the frequency distribution based on distance to that element of the RDF vector. Thus, the element of the vector with index is given by Equation 3.

Equation 3 $$A(d_n) = \frac{1}{2}\sum_{i,j,i\neq j} p_i p_j e^{-B(d_n-d_{ij})^2}$$

In this equation, and are the atomic properties, is a temperature-dependent constant that is a measure of smoothness, and is the mathematical constant [10].

Artificial Neural Networks and QSAR.

Machine learning techniques use the descriptor function values derived from the molecules in the dataset to generate a QSAR, which can then be used to predict molecular bioactivity. The machine learning technique used in this project was the Artificial Neural Network (ANN), whose structure is described in Figure 1. The ANN contains an input layer (a) where descriptors of a molecule are provided. The hidden layer (b) finds the complex, nonlinear relationships between the structure encoded by these descriptors and the biological activity of the molecules using a QSAR. The output layer (c) outputs the predicted activity. The ANN trains on 90% of the dataset to generate the QSAR model. The ANN is then tested on the independent 10% portion of the dataset to determine its prediction accuracy. The QSAR model is developed when the ANN is being trained by relating biological activity to the chemical descriptors of the initial 90% of the dataset. For this reason, the mathematical description of the molecules is pivotal, as it can significantly affect the quality if the resulting QSAR model. This then affects the functional predictions of the molecules of interest. An ANN was trained with and without the novel paired descriptors for comparison. [11–13]

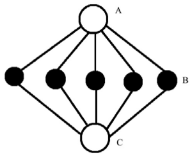


Figure 1. Schematic of Artificial Neural Network. (a) The input layer consists of the information given to the ANN, such as descriptor vectors of a portion of the dataset. (b) The hidden layer is the layer that utilizes QSAR to determine activity. (c) The output layer combines the nodes in the hidden layer to predict a final activity value of the molecule [11-13].

Descriptor Selection.

In order to minimize the noise when building the model, it is necessary to choose an optimal subset of descriptors. This reduces the total number of inputs, and increases the efficiency of the QSAR model. The method used to select the best set of descriptors was Information Gain (IG), which provides a score for each descriptor based on statistical evaluations. IG measures the order of randomness in a descriptor function compared to the rest of the descriptor category columns. Information Gain is given by Equation 4.

Equation 4 $$IG = -\sum_{i=1}^{n} x_i \log_{10} x_i$$

Here, is the *i*th feature of the descriptors summed over all compounds in the active dataset.

RESULTS.

ROC Curves.

A ROC curve is a graphical analysis tool that measures the accuracy and quality of a model. On the x-axis is the false positive rate, which is the rate of compounds that were predicted by the model to be active, but were actually inactive. On the y-axis is the true positive rate, which is the rate of molecules that were correctly predicted by the model to be active. The line on this graph represents a model that is untrained, such as a model that randomly predicts the activity of a molecule given its structure.

Quality Measures.

a) Area Under the Curve (AUC)

A more accurate model will have a ROC curve that is above the line , so that it may always have a higher true positive rate than the false positive rate. A quantifiable measure of the ROC curve's quality is the integral of this graph, or the Area under the Curve (AUC). The AUC quantifies the probability that the model will predict an active compound to have a higher bioactivity level than an inactive compound. This parameter ranges from 0 to 1, with 0 being the least accurate model, 0.5 being the random predictor, and 1 being a perfect model.

b) Enrichment

The second measure of model quality and accuracy derived from the ROC curve is known as enrichment. Enrichment is given by Equation 5.

Equation 5 $$E = \frac{\frac{TP}{TP+FP}}{\frac{P}{P+N}}$$

In this equation, *TP* is the true positive rate, *FP* is the false positive rate, *P* is the positive rate, and *N* is the negative rate. *TP* and *FP* are derived from the predictions of the model, while *P* and *N* are derived from the original HTS experiment. Enrichment is related to the initial slope of the ROC curve; as the initial slope increases, enrichment does so as well. Enrichment is considered to be the most important measure of model quality because it focuses on the region of the ROC curve with very low false positive rate, and in this region, the model is most accurate and most relevant for computational drug discovery campaigns.

QSAR Model Performance.

Figure 2 below shows the logarithmic ROC curves generated for the AID 891 dataset.

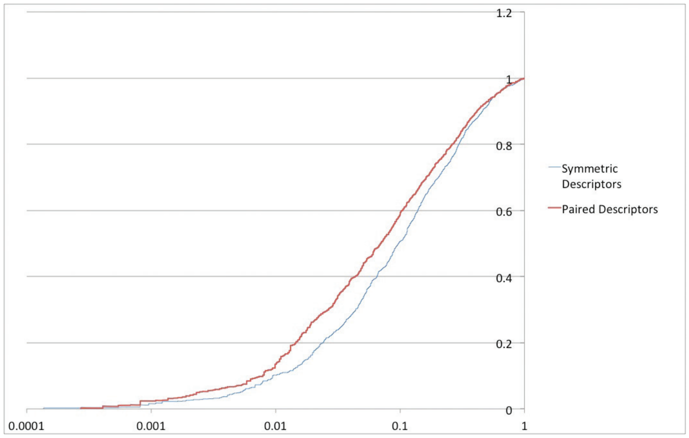


Figure 2. Logarithmic scale ROC curve for AID 891 Database

Table 1 displays the values and percent changes in AUC and Enrichment.

Table 1. Objective Function Values

Dataset ID	AUC			Enrichment		
	symmetric	paired	%	symmetric	paired	%
891	0.826	0.85	2.79	3.94	4.51	14.3

The results from Figure 2 and Table 1 are measures of the quality and accuracy of the model, and provide critical information regarding the descriptive power of paired property descriptors. The AUC and Enrichment describe the ability of the model to correctly predict the activity of the molecules. The relative change in these objective measures of quality between the unpaired descriptors and the paired property descriptors introduced in this paper reflects upon the efficiency and power of the new descriptor classes. We can see that there was a percent increase of 2.79% in AUC, and 14.3% in Enrichment.

DISCUSSION.

The goal of this project is to improve the description of molecules using a novel implementation of paired property descriptors that could improve QSAR modeling for drug discovery. The results indicate that paired descriptors improve the mathematical encoding of the constitution and configuration of molecules, at least for the sample dataset chosen.

To understand the results of the ROC curves, it is important to consider the improvements that the paired property implementation provides. The new implementation allows for the encoding of the correlations between two different properties of atoms within a particular distance interval and the biological activity. This permits a more detailed encoding of molecular structure, so that the ANN could more readily learn the structure activity relationship. Similarly, a more detailed encoding of molecular structures provided by the paired property descriptors led to a more effective model that explained a larger amount of the given data, as can be seen through AUC and Enrichment data increases of 2.79% and 14.3%, respectively.

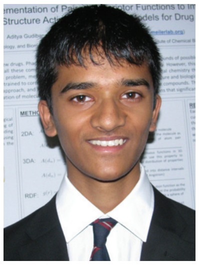
The results from ROC curve analysis indicate a significant improvement in both AUC as well as enrichment when the paired property descriptors were introduced to the Artificial Neural Network. The ANN was more likely to provide an active molecule with a higher score than an inactive molecule because the AUC and Enrichment of the new model were higher. Changes in objective function results surpass many similar studies in computational biology [1–4], [9], [14]. These improvements suggest promising applications of paired property descriptors in the prediction of novel active molecules for treatment of diseases. The novel descriptors introduced in this study can identify drug hits with greater quality and accuracy, speeding up hit-to-lead optimization. Future experiments plan to extend testing of paired property descriptors to more datasets to understand the applicability of these novel descriptors to more diverse molecular libraries. It is hoped that paired property descriptors will increase the ability of current models to predict drug hits, and speed up the process of drug discovery for the treatment of diseases.

ACKNOWLEDGMENTS.

We would like to acknowledge members of the Meiler lab including Sandeep Kothiwale, Mariusz Butkiewicz, and Jeffrey Mendenhall. We would also like to thank Jonathan Creamer and the School for Science and Math for their support. The project described was supported by Award Number R25RR024261 from the National Center For Research Resources. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center For Research Resources or the National Institutes of Health.

REFERENCES.

1. R. Mueller, A. L. Rodriguez, E. S. Dawson, *et al.*, *ACS Chem Neurosci.* 1, 288 (2010).
2. M. Butkiewicz, R. Mueller, D. Selic, *et al.*, *Neural Networks.* 255 (2009).
3. R. Mueller, E. S. Dawson, J. Meiler, *et al.*, *Chem Med Chem.* 7, 406 (2012).
4. G. Sliwoski, E. W. Lowe, M. Butkiewicz, *et al.*, *Molecules.* 17, 9971 (2012).
5. "The PubChem Project." [Online]. *NIH* (2012).
6. B. F. Jensen, C. Vind, S. B. Padkjaer, *et al.*, *J Med Chem.* 50, 501 (2007).
7. U. M. Zanger, S. Raimundo, M. Eichelbaum, *Naunyn-Schmiedeberg's Arch. Pharmacol.* 369, 23 (2004).
8. G. Franzese, a. Fierro, a. De Candia, *et al.*, *PHYA.* 257, 376 (1998).
9. M. Butkiewicz, E. W. Lowe, J. Meiler, *Proc IEEE CIBCB.*, 329–334 (2012).
10. A. Code, "ADRIANA.Code," *Program*, (2011).
11. I. a Basheer M. Hajmeer, *J Microbiol Methods*, vol. 43, no. 1, pp. 3–31, Dec. 2000.
12. M. Moliner, J. M. Serra, a. Corma, E. Argente, S. Valero, and V. Botti, *Microp M M.* 78, 73 (2005).
13. G. Zhang, B. E. Patuwo, and M. Y. Hu, *I J Forecasting.* 14, 35 (1998).
14. E. W. Lowe, A. Ferrebee, A. L. Rodriguez, *et al.*, *Bioorg. Med. Chem. Lett.* 20, 5922 (2010).



Aditya M. Gudibanda is a student at Hume-Fogg Academic Magnet High School in Nashville, Tennessee, and enrolled in the School for Science and Math at Vanderbilt.